

The VC-Dimension of Similarity Hypothesis Spaces

Mark Herbster, Paul Rubenstein, James Townsend

Department of Computer Science
University College London
Gower Street, London WC1E 6BT, England, UK
{m.herbster, paul.rubenstein.14, james.townsend.14}@ucl.ac.uk

February 26, 2015

Abstract

Given a set X and a function $h : X \rightarrow \{0, 1\}$ which labels each element of X with either 0 or 1, we may define a function $h^{(s)}$ to measure the *similarity* of pairs of points in X according to h . Specifically, for $h \in \{0, 1\}^X$ we define $h^{(s)} \in \{0, 1\}^{X \times X}$ by $h^{(s)}(w, x) := \mathbb{1}[h(w) = h(x)]$. This idea can be extended to a set of functions, or *hypothesis space* $\mathcal{H} \subseteq \{0, 1\}^X$ by defining a *similarity hypothesis space* $\mathcal{H}^{(s)} := \{h^{(s)} : h \in \mathcal{H}\}$. We show that $\text{VC-DIMENSION}(\mathcal{H}^{(s)}) \in \Theta(\text{VC-DIMENSION}(\mathcal{H}))$.

1 Introduction

Consider the problem of learning from examples. We may learn by receiving *class* labels as feedback: ‘this is a **dog**’, ‘that is a **wolf**’, ‘there is a **cat**’, etc. We may also learn by receiving *similarity* labels: ‘these are the **same**’, ‘those are **different**’ and so forth. In this note we study the problem of learning with similarity versus class labels. Our approach is to use the *VC-dimension* [VC71] to study the fundamental difficulty of this learning task.

In the supervised learning model we are given a training set of patterns and associated labels. The goal is then to find a hypothesis function that maps patterns to labels that will predict with few errors on future data (*small generalization error*). A classic approach to this problem is empirical risk minimisation. Here the procedure is to choose a hypothesis from a set of hypothesis functions (*hypothesis space*) that ‘fits’ the data as closely as possible. If the hypothesis is from a hypothesis space with small VC-dimension and fits the data well then we are likely to predict well on future data [VC71, BEHW89]. The number of examples required to have small generalisation error with high probability is called the *sample complexity*. In the *uniform learnability* model the VC-dimension gives a nearly matching upper and lower bound on the sample complexity [BEHW89, EHKV89]. In Theorem 1 we demonstrate that the VC-dimension of a hypothesis space with respect to similarity-labels is proportionally bounded by the VC-dimension with respect to class-labels indicating that the sample complexities within the two feedback

settings are comparable. That is, the fundamental difficulties of the two learning tasks are comparable.

Related work

We are motivated by the results of [GHP13]. Here the authors considered the problem of similarity prediction in the online mistake bound model [Lit88]. In [GHP13, Theorem 1] it was found that given a basic algorithm for class-label prediction with a mistake bound there exists an algorithm for similarity-label prediction with a mistake bound which was larger by no more than a constant factor. In this work we find an analogous result in terms of the VC-dimension.

2 The VC-dimension of similarity hypothesis spaces

A hypothesis space $\mathcal{H} \subseteq \{0,1\}^X$ is a set of functions from some set of patterns X to the set of labels $Y = \{0,1\}$ in the two-class setting. The *restriction* of a function $h \in \{0,1\}^X$ to a subset $X' \subseteq X$ is the function $h|_{X'} \in \{0,1\}^{X'}$ with $h|_{X'}(x) := h(x)$ for each $x \in X'$. Analogously, one can define the restriction of a hypothesis space as $\mathcal{H}|_{X'} := \{h|_{X'} : h \in \mathcal{H}\}$.

A subset $X' \subseteq X$ is said to be *shattered* by \mathcal{H} if $\mathcal{H}|_{X'} = \{0,1\}^{X'}$, that is if the restriction contains *all possible* functions from X' to $\{0,1\}$. The VC-dimension [VC71] of a hypothesis space $\mathcal{H} \subseteq \{0,1\}^X$, denoted $d(\mathcal{H})$, is the size of the largest subset of X which is shattered by \mathcal{H} , that is

$$d(\mathcal{H}) := \max_{X' \subseteq X} \{|X'| : \mathcal{H}|_{X'} = \{0,1\}^{X'}\}.$$

Sauer's lemma [VC71, Sau72, She72], which gives a lower bound for the VC-dimension of a hypothesis space, will be used for proving our main result. It states that for a hypothesis space $\mathcal{H} \subseteq \{0,1\}^X$, if

$$|\mathcal{H}| > \sum_{k=0}^{m-1} \binom{|X|}{k} \tag{1}$$

then $d(\mathcal{H}) \geq m$.

Given a function $h : X \rightarrow \{0,1\}$, we may define a function $h^{(s)}$ to measure the *similarity* of pairs of points in X according to h . Specifically, for $h \in \{0,1\}^X$ we define $h^{(s)} \in \{0,1\}^{X \times X}$ by $h^{(s)}(w, x) := \mathbb{1}[h(w) = h(x)]$, where $\mathbb{1}$ is the indicator function. This idea can be extended to a hypothesis space \mathcal{H} by defining the *similarity hypothesis space* $\mathcal{H}^{(s)} := \{h^{(s)} : h \in \mathcal{H}\}$. We now give our central result,

Theorem 1. *Given a hypothesis space $\mathcal{H} \subseteq \{0,1\}^X$,*

$$d(\mathcal{H}) - 1 \leq d(\mathcal{H}^{(s)}) \leq \delta d(\mathcal{H}),$$

with $\delta = 4.55$.

Proof. For the left hand inequality, let $n := d(\mathcal{H})$ and pick a set $T = \{x_1, x_2, \dots, x_n\}$ of size n which is shattered by \mathcal{H} . Then let $T' = \{(x_1, x_2), (x_2, x_3), \dots, (x_{n-1}, x_n)\}$. To demonstrate that T' is shattered by $\mathcal{H}^{(s)}$, let $g \in \{0, 1\}^{T'}$ be any mapping from T' to $\{0, 1\}$. Then since T is shattered by \mathcal{H} we may find a map $h \in \mathcal{H}$ with $h(x_1) = 0$ and

$$h(x_{i+1}) = \begin{cases} h(x_i) & \text{if } g(x_i, x_{i+1}) = 1 \\ 1 - h(x_i) & \text{if } g(x_i, x_{i+1}) = 0 \end{cases}$$

for $i = 1, \dots, n-1$. Observe that $g = h^{(s)}|_{T'}$. Since g was chosen arbitrarily, we may conclude that T' is indeed shattered by $\mathcal{H}^{(s)}$, and therefore $d(\mathcal{H}^{(s)}) \geq |T'| = d(\mathcal{H}) - 1$.

For the right hand inequality, first let $M := d(\mathcal{H}^{(s)})$ and then pick a set $U = \{(w_1, x_1), (w_2, x_2), \dots, (w_M, x_M)\}$ of size M in $X \times X$ which is shattered by $\mathcal{H}^{(s)}$. Let $V = \{w_1, w_2, \dots, w_M, x_1, x_2, \dots, x_M\}$ and note that $|\mathcal{H}|_V \geq |\mathcal{H}^{(s)}|_U = 2^M$. This is because any two maps h and g which agree on V will induce maps $h^{(s)}$ and $g^{(s)}$ which agree on U , so $\mathcal{H}^{(s)}|_U$ cannot possibly contain more maps than $\mathcal{H}|_V$. Using this fact, and applying Sauer's Lemma (see (1)) to $\mathcal{H}|_V$, we see that if

$$2^M > \sum_{k=0}^{m-1} \binom{|V|}{k}$$

then $d(\mathcal{H}) \geq d(\mathcal{H}|_V) \geq m$.

Now note the following inequality (see e.g., [FG06, Lemma 16.19]), which bounds a sum of binomial coefficients:

$$\sum_{i=0}^{\lfloor \epsilon n \rfloor} \binom{n}{i} \leq 2^{H(\epsilon)n} \quad (0 < \epsilon < 1/2), \quad (2)$$

where $H(\epsilon) := \epsilon \log_2 \frac{1}{\epsilon} + (1 - \epsilon) \log_2 \frac{1}{1-\epsilon}$ denotes the *binary entropy* function. If we set $m = 1 + \lfloor 2\epsilon M \rfloor$ for some $\epsilon < \frac{1}{2}$ such that $H(\epsilon) < \frac{1}{2}$, we have

$$\sum_{k=0}^{m-1} \binom{|V|}{k} = \sum_{k=0}^{\lfloor 2\epsilon M \rfloor} \binom{|V|}{k} \leq \sum_{k=0}^{\lfloor 2\epsilon M \rfloor} \binom{2M}{k} \leq 2^{2MH(\epsilon)} < 2^M$$

using (2) and that $|V| \leq 2M$ from the definition of V . Thus Sauer's lemma can be applied with the above value of m and hence

$$d(\mathcal{H}) \geq 1 + \lfloor 2\epsilon M \rfloor \geq 2\epsilon M = 2\epsilon d(\mathcal{H}^{(s)}),$$

as long as $H(\epsilon) < 1/2$. Observe that $\epsilon = .11$ satisfies this condition and thus we have that

$$d(\mathcal{H}^{(s)}) \leq 4.55d(\mathcal{H}).$$

□

3 Discussion

In the following, we give a family of examples where the VC-dimension of the similarity hypothesis space is exactly twice that of the original space. We use the following notation for the set of the first n natural numbers $[n] := \{1, 2, \dots, n\}$.

Example 2. For the hypothesis space of k -sparse vectors, $\mathcal{H}_k := \{h \in \{0, 1\}^{[n]} : \sum_{i=1}^n h(i) \leq k\}$,

$$d(\mathcal{H}_k) = k \text{ and } d(\mathcal{H}_k^{(s)}) = 2k,$$

provided that $n \geq 2k + 1$.

Proof. Let $X := [n]$. Firstly note that $d(\mathcal{H}_k) \geq k$, since any subset $T \subseteq X$ with $|T| \leq k$ is shattered by \mathcal{H}_k . If $T' \subseteq X$ with $|T'| > k$ then T' cannot possibly be shattered by \mathcal{H}_k since there is no element in \mathcal{H}_k that labels all elements of T' as 1. Therefore $d(\mathcal{H}_k) = k$.

To see that $d(\mathcal{H}_k^{(s)}) \geq 2k$, let $U = \{(x_1, x_2), (x_2, x_3), \dots, (x_{2k}, x_{2k+1})\}$ for any distinct elements $x_1, x_2, \dots, x_{2k+1} \in X$ and note that $|U| = 2k$. To show that U is shattered by $\mathcal{H}_k^{(s)}$, let $g \in \{0, 1\}^U$ be any function from U to $\{0, 1\}$. We need to find an $h \in \mathcal{H}_k$ such that $g = h^{(s)}|_U$. Two functions in $\{0, 1\}^X$ which satisfy the condition $g = h^{(s)}|_U$ are h_0 and h_1 defined by $h_0(x_1) = 0$, $h_1(x_1) = 1$ and

$$\begin{aligned} h_j(x_{i+1}) &= \begin{cases} h_j(x_i) & \text{if } g(x_i, x_{i+1}) = 1 \\ 1 - h_j(x_i) & \text{if } g(x_i, x_{i+1}) = 0 \end{cases} \\ h_j(x) &= 0 \quad \forall x \notin \{x_1, x_2, \dots, x_{2k+1}\} \end{aligned}$$

for $i = 1, \dots, 2k$ and $j = 0, 1$. Observe that by construction, $h_0(x_i) + h_1(x_i) = 1$ for each $i = 1, \dots, 2k + 1$ and therefore $\sum_{i=1}^{2k+1} h_0(x_i) + \sum_{i=1}^{2k+1} h_1(x_i) = \sum_{i=1}^{2k+1} [h_0(x_i) + h_1(x_i)] = 2k + 1$. This means that we must have $\sum_{i=1}^{2k+1} h_j(x_i) \leq k$ for some j and hence $h_j \in \mathcal{H}_k$ with $h_j^{(s)}|_U = g$. This proves that $d(\mathcal{H}_k^{(s)}) \geq 2k$.

Now suppose, for a contradiction, that $d(\mathcal{H}_k^{(s)}) > 2k$. Then there is some set $E = \{(u_1, v_1), (u_2, v_2), \dots, (u_{2k+1}, v_{2k+1})\} \subseteq X \times X$ of size $2k + 1$ which is shattered by $\mathcal{H}^{(s)}$. Let $V := \{u_1, u_2, \dots, u_{2k+1}, v_1, v_2, \dots, v_{2k+1}\}$ (note that in general we do not necessarily have that $|V| = 4k + 2$ since the u_i and v_i need not all be distinct).

Let G be the graph with vertex set V and edge set E . Observe that elements of \mathcal{H}_k correspond to $\{0, 1\}$ -labellings of V and that elements of $\mathcal{H}_k^{(s)}$ correspond to $\{0, 1\}$ -labellings of E . Since E is shattered by $\mathcal{H}_k^{(s)}$, every labelling of E is realisable as the induced map $h^{(s)}$ of some $h \in \mathcal{H}_k$.

Note that G cannot contain a cycle since there is no labelling of V which could induce a similarity labelling on a cycle in which exactly one edge is labelled 0 and the rest are labelled 1*. So the graph is a union of trees, also known as a ‘forest’. Note that in

*Indeed, under any such labelling of E any two vertices in the cycle are connected by two paths, one path containing exactly zero edges labelled with a 0 (implying that the two vertices are labelled the same) and one path containing exactly one edge labelled with a 0 (implying that the two vertices are labelled differently).

general the number of vertices in a forest is $|V| = |E| + r$, where $|E|$ is the number of edges and r is the number of trees in the forest. In this case we have $|V| = 2k + 1 + r$.

Now choose a labelling g , which labels the vertices of each connected component (tree) in G according to the following rule: for each connected component C in G , label $\lfloor \frac{|C|}{2} \rfloor$ vertices $v \in C$ with a 1 and the remaining $\lceil \frac{|C|}{2} \rceil$ with a 0. Note that $g \notin \mathcal{H}_k$ since

$$\sum_{v \in V} g(v) = \sum_C \sum_{v \in C} g(v) = \sum_C \left\lfloor \frac{|C|}{2} \right\rfloor \geq \sum_C \frac{|C| - 1}{2} = \frac{|V| - r}{2} = k + \frac{1}{2} > k.$$

Consider the edge labelling $g^{(s)}|_E$. Since E is shattered by $\mathcal{H}_k^{(s)}$, there must be some $h \in \mathcal{H}_k$ such that $h^{(s)}|_E = g^{(s)}|_E$. But this is not possible, for if it were, then in order for $h^{(s)}$ to agree with $g^{(s)}$ we would need $h|_C = g|_C$ or $h|_C = 1 - g|_C$ for each connected component C in G . Swapping the labellings between 0 and 1 on one or more of the connected components can only increase the number of 1 labellings and thus

$$\sum_{v \in V} h(v) \geq \sum_{v \in V} g(v) > k$$

so h cannot be in \mathcal{H}_k . Thus we have found a labelling of E , namely $g^{(s)}|_E$, which cannot be in $\mathcal{H}_k^{(s)}$. But this is a contradiction of our initial assumption that E was shattered by $\mathcal{H}_k^{(s)}$. So we have proved that our assumption must have been incorrect and therefore $d(\mathcal{H}_k^{(s)}) = 2k$. □

In Theorem 1, the lower bound $d(\mathcal{H}) - 1 \leq d(\mathcal{H}^{(s)})$ is tight, for example when $\mathcal{H} = \{0, 1\}^{[n]}$. However, observe that in Example 2, the hypothesis space of k -sparse vectors, the similarity space “expands” only by a factor of 2, which is less than the factor $\delta = 4.55$ of Theorem 1. We leave as a conjecture that the upper bound in Theorem 1 can be improved to a factor of two.

Acknowledgements. We would like to thank Shai Ben-David, Ruth Uerner and Fabio Vitale for valuable discussions. In particular we would like thank Ruth Uerner for proving an initial motivating upper bound of $d(\mathcal{H}^{(s)}) \leq 2d(\mathcal{H}) \log(2d(\mathcal{H}))$.

References

- [BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.
- [EHKV89] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989. First appeared in Proc. 1st Annu. Workshop on Comput. Learning Theory, 1988.

- [FG06] J. Flum and M. Grohe. *Parameterized Complexity Theory (Texts in Theoretical Computer Science. An EATCS Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [GHP13] Claudio Gentile, Mark Herbster, and Stephen Pasteris. Online similarity prediction of networked data from known and unknown graphs. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Proceedings*, pages 662–695. JMLR.org, 2013.
- [Lit88] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, April 1988.
- [Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972. cited By 255.
- [She72] Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41(1):247–261, 1972.
- [VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.